Weekly Report

October 14, 2018

1 Work

1. MemoryGAN的论文，在初稿的基础上对文字进行了梳理扩展，加入了当前算法的结果图片，还需要生成其他算法的图片进行对比。
2. 低光照图片增强的论文，引入了1）GAN，但是效果有所下降；2）multi stage，多层次对最终的结果略有提升。
3. 工作时长：工作日每天10个小时，周末共8个小时，共58个小时。

1.1 工作进度

Table 1: 工作进度

项目	进度	截止时间
DRGraph	需要对程序做一些修改	12.30
降维	论文修订	
专利	完成撰写，等待律师回复	
CVPR投稿（Memory GAN）	修订初稿中	11.1
CVPR投稿（See in the dark）	正在探索网络结构	11.15

2 Paper Reading

2.1 MSR-net:Low-light Image Enhancement Using Deep Convolutional Network

基于Single-scale Retinex(SSR)，使用神经网络来拟合物理模型

KVM-GAN: Key-Value Memory Generative Adversarial Networks for Text-to-image Synthesis

Anonymous CVPR submission

Paper ID ****

Abstract

Synthesizing images from text descriptions is known to be an important task recently. Generative Adversarial Networks (GANs) are widely employed to generate images based on a sentence of description.However, current GANs are still far from satisfaction. Most text-to-image synthesis methods simply combine the text information as a condition.

We propose the Key-Value Memory Generative Adversarial Networks (KVM-GAN) with a memory component to utilize text information more effectively. In particular, the key-value memory is written dynamically conditioned on word information and a initial generated image. Then, the initial image is enhanced with fine-grained details via nontrivial transforms between key and value memories. Our proposed method is evaluated on two datasets, i.e., the Caltech-UCSD Birds 200 (CUB) dataset and the Microsoft Common Objects in Context (COCO) dataset. We compare the KVM-GAN model with existing models using the Inception Score (IS), Fréchet Inception Distance (FID) and R-precision. Experimental results demonstrate that our KVM-GAN model achieves appealing performance compared with the state-of-the-art, both quantitatively and qualitatively.

1. Introduction

The last few years have seen a remarkable growth in the use of Generative Adversarial Networks (GANs) [3] for image generation. Image synthesis based on the text description has been an important problem in artificial intelligence nowadays. Fully understanding the relationship between visual contents and natural languages is an essential step towards artificial intelligent, e.g., image search and summarization.

Recently, most text-to-image synthesis methods employ Generative Adversarial Networks (GANs) to generate images based on text descriptions (see Figure 1). To generate images that are well conditioned on input text descriptions, a joint representation is learned to bridge the gap between visual contents and natural languages. Most existing text-to-image synthesis methods [20] simply combine text and image feature vectors. AttnGAN [24] is conditioned on the fine-grained word level information (e.g., word-context features) via attention mechanism. However, we contend that it is more powerful to dynamically reason over regions of the image and word level information.

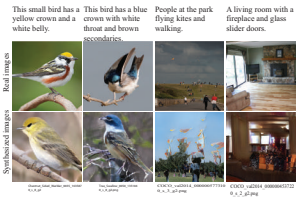

Figure 1. Example results of text-to-image synthesis by our KVM-GAN model.

Memory network [26] offers a new architecture to recurrently read from text memory multiple times. Then, the textual features are combined together to generate a response. The Key-Value memory network [11] further allows encoding prior knowledge for the considered task and leveraging more complex transforms for memory operations. It gives the model greater flexibility for encoding and reason knowledge sources.

Inspired by the recent success of memory network, we propose a novel the Key-Value Memory Generative Adversarial Networks (KVM-GAN) which integrates the memory component in the context of AttnGAN [24] to enable a dynamic reasoning over images and text descriptions. The KVM-GAN writes word-level information as well as image information into a key-value structured memory for text-to-image synthesis after generating an initial image according

Figure 1: #1

2.2 Multimodal Unsupervised Image-to-Image Translation

文章想要解决image-to-image中many to many的问题。文章认为一张图片的特征（隐变量）包含了两个信息，content和style。content就是图片的基本内容，而style就是我们想要变化的风格。我们可以通过训练网络把content和style分别抽取出来，然后改变style的数值，从而实现到另一种类型图片的转换。

2.3 FashionNet: Personalized Outfit Recommendation with Deep Neural Network

set recommendation problem是指要推荐一些列搭配的物体。本文把这个问题分解为1）物体的特征抽取；2）两种搭配的距离比较。作者使用神经网络学习几个物体搭配在一起的特征，然后使用rank loss，搭配更和谐的需要取得更高的rank。

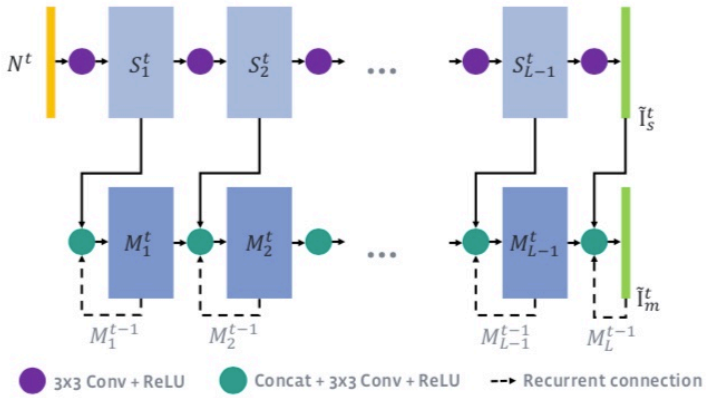

Figure 2: #2

(a) User 1 (b) User 2 (c) User 3

Figure 3: #3

2.4 Diverse Image-to-Image Translation via Disentangled Representations

同样类似的隐变量分解的方法，只不过又增加了一些loss。

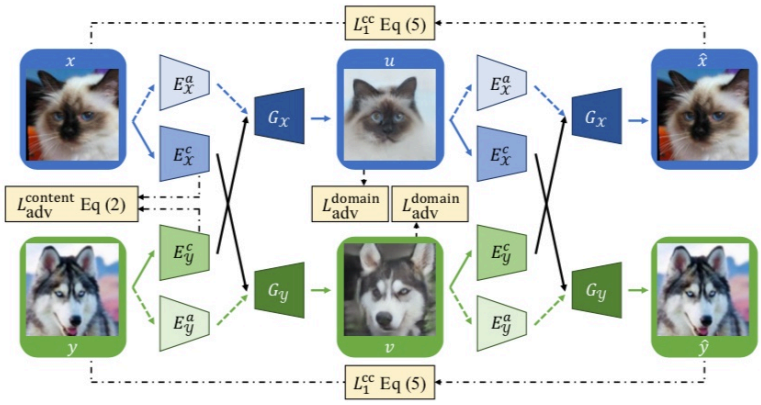

Figure 4: #4

KVM-GAN: Key-Value Memory Generative Adversarial Networks for Text-to-image Synthesis

Anonymous CVPR submission

Paper ID ****

Abstract

Synthesizing images from text descriptions is known to be an important task recently. Generative Adversarial Networks (GANs) are widely employed to generate images based on a sentence of description.However, current GANs are still far from satisfaction. Most text-to-image synthesis methods simply combine the text information as a condition. We propose the Key-Value Memory Generative Adversarial Networks (KVM-GAN) with a memory component to utilize text information more effectively. In particular, the key-value memory is written dynamically conditioned on word information and a initial generated image. Then, the initial image is enhanced with fine-grained details via nontrivial transforms between key and value memories. Our proposed method is evaluated on two datasets, i.e., the Caltech-UCSD Birds 200 (CUB) dataset and the Microsoft Common Objects in Context (COCO) dataset. We compare the KVM-GAN model with existing models using the Inception Score (IS), Fréchet Inception Distance (FID) and R-precision. Experimental results demonstrate that our KVM-GAN model achieves appealing performance compared with the state-of-the-art, both quantitatively and qualitatively.

1. Introduction

The last few years have seen a remarkable growth in the use of Generative Adversarial Networks (GANs) [3] for image generation. Image synthesis based on the text description has been an important problem in artificial intelligence nowadays. Fully understanding the relationship between visual contents and natural languages is an essential step towards artificial intelligent, e.g., image search and summarization.

Recently, most text-to-image synthesis methods employ Generative Adversarial Networks (GANs) to generate images based on text descriptions (see Figure 1). To generate images that are well conditioned on input text descriptions, a joint representation is learned to bridge the gap between visual contents and natural languages. Most existing text-to-image synthesis methods [20] simply combine text and image feature vectors. AttnGAN [24] is conditioned on the fine-grained word level information (e.g., word-context features) via attention mechanism. However, we contend that it is more powerful to dynamically reason over regions of the image and word level information.

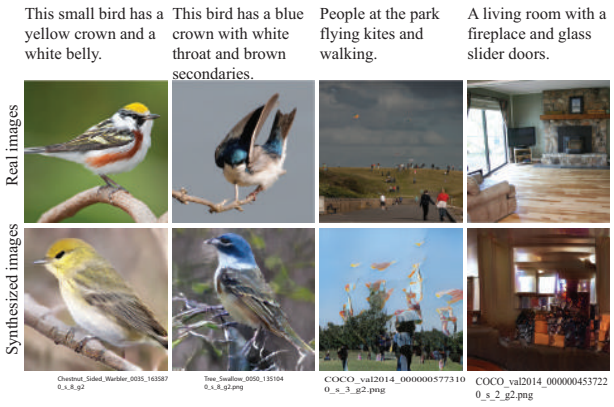

Figure 1. Example results of text-to-image synthesis by our KVM-GAN model.

Memory network [26] offers a new architecture to recurrently read from text memory multiple times. Then, the textual features are combined together to generate a response. The Key-Value memory network [11] further allows encoding prior knowledge for the considered task and leveraging more complex transforms for memory operations. It gives the model greater flexibility for encoding and reason knowledge sources.

Inspired by the recent success of memory network, we propose a novel the Key-Value Memory Generative Adversarial Networks (KVM-GAN) which integrates the memory component in the context of AttnGAN [24] to enable a dynamic reasoning over images and text descriptions. The KVM-GAN writes word-level information as well as image information into a key-value structured memory for text-to-image synthesis after generating an initial image according

to the whole input sentence. The memory writing operation reshapes the word embedding by focusing on the pixels which are relevant to the word. Then, the memory component is employed to combine the initial image and predict a more realistic image, where the key memory is used to address the relevant text information with respect to the initial image, and the value memory is used to return the corresponding memory. At last, a adaptive gating mechanism is employed to control the combination of memory and image features. We conduct experiments to evaluate the KVM-GAN model with existing models using the Inception Score (IS), Fréchet Inception Distance (FID) and R-precision. The experiments on the CUB and COCO datasets demonstrate that the MemoryGAN outperforms the previous image-to-text synthesis methods. Specifically, our model improve IS from 4.36 to 4.80 and decrease the FID from 23.98 to 16.33 on the CUB dataset. The R-precision, which measure the matching degree between input text description and the generated image, is improved by 4% on two datasets. Qualitative evaluation shows that our generative model generates more photo-realistic high-resulution images. In summary, we make the following contributions.

- We propose the MemoryGAN with a memory component to conditionally improve the image quality based on the text memory.
- We evaluate the MemoryGAN on the CUB and COCO datasets, showing competitive performance with other text-to-image models.

The rest of this paper is organized as follows. In section 2, we review the related work on generative adversarial networks and memory networks. Section 3 describe our KVM-GAN model incorporating a memory component.In Section 4, we present an experimental evaluation on the CUB and COCO datasets. Finally, we draw conclusions in Section 5.

2. Related Work

Generative Adversarial Networks. With the recent successes of Variational Autoencoders (VAEs) [7] and GANs [3], a large number of methods have been proposed to handle image generation task. Recently, GANs have received great attention because of their capability to produce sharp images. Apart from conditional generation models (e.g., cGAN [12] and AC-GAN [15]) which generate images according to the given class label, generating images based on the text descriptions gains interest in the research community nowadays.

The text-to-image synthesis problem is decomposed by Reed et al.[20] into two sub-problems: first, the joint embedding is learned to capture the relations between natural language and real-world images; second, a deep convolutional generative adversarial network [18] is trained to synthesize a compelling image. Dong et al. [2] adopted the pair-wise ranking loss [8] to project both images and texts into a joint embedding space. Since previous generative model fail to add the location information, Reed et al. proposed GAWWN [19] to encode localization constraints. StackGAN [28] and StackGAN++ [29] generated photo-realistic high-resolution images with two stages. In addition, StackGAN employed a novel Conditioning Augmentation technique to increase the training dataset and stabilize the training. To diversify the generated images, the discriminator of TAC-GAN [1] not only distinguish real images from synthetic images, but also classifier synthetic images into true classes. Similar to TAC-GAN, PPGN [14] included a conditional network to guided the generator to synthesize images conditioned on a caption or a class. Yuan et al. [27] employed symmetrical distillation networks to minimize the multi-level difference between real and synthetic images. Conditioning on the global sentence vector may results in lower quality images, AttnGAN [24] first generated low resolution images and then refined the image to high resolution by using attention mechanism over the image and word features. Besides, AttnGAN also proposed an attention based model to map images and words to a common semantic space. However, most text-to-image synthesis methods simply use the text information as the condition. For instance, directly concatenating the image and text feature may resulted in neglecting either image or text information.

Memory Networks. Recently, memory network [26, 17, 4] provides a new architecture to reason answers from memories more effectively using explicit storage and a notion of attention. Memory network first writes information into a external memory and then reads contents from memory slots according to a relevance probability. Weston et al. [26] introduced memory network with the memory component to produce the output by searching the supporting memories one by one. End-to-end memory network [22] is a continues form of memory network, where each memory is weighted according to the inner product between the memory and the input. Liu et al. [10] further employed highway network to control the utilization of memory information based on the current input. To understand the unstructured documents, the key-value memory network [11] performed reasoning by utilizing different encodings for on the key-value structured memory. The key memory are used to infer the weight of the corresponding value memory when predicting the final answer. Inspired by the recent success of memory network, we introduce KVM-GAN, a novel network architecture to update the key-value structured memory and enhance synthetic images with fine-grained details via nontriv-ial transforms between key and value memories.

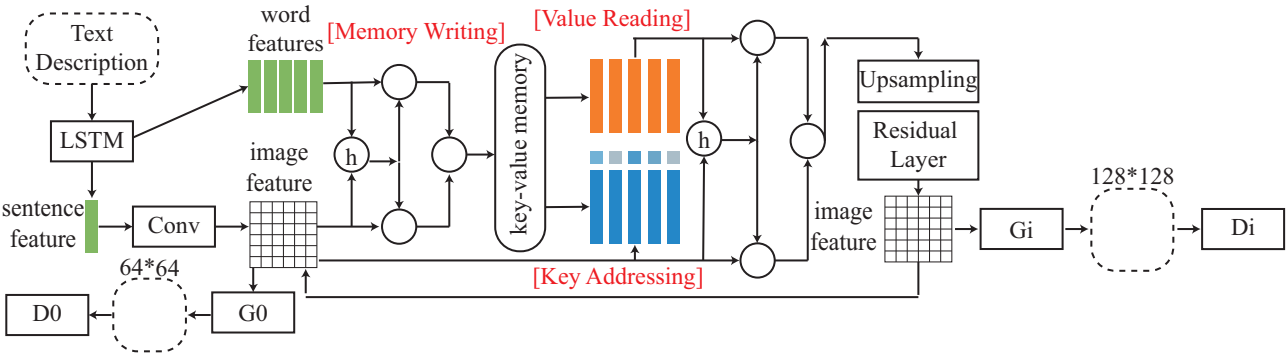

Figure 2. The MemoryGAN architecture for text-to-image synthesis.

3. Key-Value Memory GAN

3.1. Overview

The KVM-GAN model is based on the architecture of the AttnGAN [24] which significantly improves the image quality. Figure 2 illustrate the overview of KVM-GAN model architecture, including the key-value structured memory and the mutual-attention mechanism. Our approach consists of three stages: text embedding, initial image generation, and conditional image refinement.

Text Embedding. The target text description are encoded into a semantic representation with the deep attentional multimodal similarity model (DAMSM) introduced by AttnGAN [24]. The text encoder is a pre-trained bidirectional LSTM encoder, which transforms the input text description into some internal representation, i.e., a sentence feature and several word features. Each word feature w_i corresponds to the two hidden states of two directions. The global sentence feature s is generated by concatenating the last hidden states of two directions.

Initial Image Generation In this stage, we generate a low-resolution image according to the sentence feature. First, the sentence feature is fed into a conditioning augmentation network, which resamples a vector from an independent Gaussian distribution to augment training data and avoid overfitting. Then, a deep conventional generator G_0 is employed to generate a image with rough shape and few details.

Conditional Image Refinement. To generate a photo-realistic image, we refine low-resolution images into high-resolution images with more details. A naive way to enhance images with word-level information is to simply concatenate image features and word-level features. We argue that it is more powerful to dynamically reason over regions of the image and words using a key-value memory component. Specifically, we first write word information combined with initial image features into memories before reasoning on them in order to refine the image. The key-value memories are designed to use key to address relevant word features for each image pixel feature and read corresponding values to generate a new image feature. Finally, we refine the image using the adaptive gating mechanism to dynamically fuse the previous image feature and the new image features. The refinement process is repeated two times to retrieve more pertinent information and generate more fine-grained details.

3.2. Image Refinement

For the given input word W and image I feature representation, we aim to learn a model to refine the image with multiple stages. The key-value structured memory provides a more effective way to utilize memory via nontrivial transforms between key and value. The key memory and value memory access process is conducted by two different neural network using $\phi_K(m_i)$ and $\phi_V(m_i)$.

The refinement operation of the KVM-GAN includes four steps:

- **Memory Writing**: Encoding prior knowledge is an important component of KVM-GAN, which enables recovering high-quality images from text. We write word information into the memory after enhancing word representation with the current image feature to improve the word representation and reduce the importance of meaningless words. Instead of directly writing word features into memory, we employ a learned gate to selectively update memory conditioned on an image. The gate g_i is computed as follows:

$$\overline{I} = \frac{1}{N}\sum_{i=1}^{N} I_i$$
$$g_i = A * w_i + B * I$$

where $\overline{I}$ is the average image feature and $A(B)$ is a matrix which embeds image or text features into the same feature space. Then, the memory m_i is written

by combining the image and word features with the gate.

$$m_i = Conv(w_i) * g_i + Conv(I) * (1 - g_i)$$

where $Conv()$ denotes a 1x1 convolution operation.

- **Key Addressing**: In this step, we retrieve relevant memories with respect to the key memory. We compute a weight of each memory slot as a similarity probability between the memory m_i and the image features I_j:

$$\alpha_{i,j} = \frac{exp(\phi_K(m_i)^T I_j)}{\sum_j exp(\phi_K(m_i)^T I_j)}$$

where $\alpha_{i,j}$ is the similarity probability between the i-th memory and the j-th image feature and $\phi_K()$ is the key memory access process.

- **Value Reading**: The final output is defined as the weighted summation of value memories according the similarity probability. The output memory representation is obtained by:

$$O_i = \sum_{j=0} \alpha_{i,j} \phi_V(m_i)$$

where $\phi_V()$ is the value memory access process.

- **Response**: After receiving the memory result, we combine the current image and the output representation to provide a new image feature. A naive approach will be simply concatenating the image features and the output representation. It will be better to let the model to decide how much information it wants to receive. We employ the adaptive gating mechanism to dynamically control the information flow and update the image feature the output representation.

$$gate = \sigma(W[O_i, I_i] + b)$$
$$I_i^{new} \leftarrow I_i * (1 - gate) + O_i * gate$$

where W and b are the parameter matrix and bias term, σ is the sigmoid function, $[\cdot, \cdot]$ denotes concatenation operation, and $gate$ is the control gate for conditional refinement.

3.3. Objective Function

The final objective function of the MemoryGAN is defined as:

$$L = \sum_i (L_{G_i} + L_{Di}) + \lambda_1 L_{CA} + \lambda_2 L_{DAMSM}$$

in which λ_1 and λ_2 are the corresponding weights.

Adversarial Loss: The adversarial loss for G_i is defined as follows:

$$L_{G_i} = \underbrace{-\frac{1}{2} E_{x \in p_{G_i}}[logD_i(x_i)]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2} E_{x \in p_{G_i}}[logD_i(X, s)]}_{\text{conditional loss}}$$

where the first term is the unconditional loss which make the generated image real as much as possible and the second term is the conditional loss which makes the image match the input sentence s. The discriminator loss also includes a unconditional loss and a conditional loss:

$$L_{Di} = \underbrace{-\frac{1}{2} E_{x \in p_{data}} logD_i(x_i) - \frac{1}{2} E_{x \in p_{G_i}} log(1 - D_i(x_i))}_{\text{unconditional loss}}$$
$$\underbrace{-\frac{1}{2} E_{x \in p_{data}} logD_i(x_i, s) - \frac{1}{2} E_{x \in p_{G_i}} log(1 - D_i(x_i, s))}_{\text{conditional loss}}$$

where the unconditional loss is designed to distinguish the generated image from real images and the conditional loss determines whether the image and the input sentence match.

Conditioning Augmentation Loss: The Conditioning Augmentation (CA) technique is proposed to augment training data and avoid overfitting by resampling input sentence vector from an independent Gaussian distribution. Thus, the CA loss is defined as the Kullback-Leibler divergence between the standard Gaussian distribution and the Gaussian distribution of training data.

$$L_{CA} = D_{KL}(N(\mu(s), \Sigma(s)) || N(0, I))$$

DAMSM Loss: The DAMSM loss is introduced to measure the matching degree between images and text descriptions. The DAMSM loss makes generated images better conditioned on the input text description. Please refer to AttnGAN by Xu et al. [24] for more details.

4. Experiment

In this section, we evaluate the MemoryGAN quantitatively and qualitatively, compared with the state-of-the-art methods [24]. We implement the proposed KVM-GAN using the open-source Python library PyTorch [16]. We conduct all experiments on a desktop PC with two NVIDIA GTX 1080 Ti and Ubuntu 16.04 installed.

Datasets. To demonstrate the capability of our proposed method for text-to-image synthesis, we conduct experiments on the Caltech-UCSD Birds 200 (CUB) [25] and the Microsoft Common Objects in Context (COCO) [9] datasets.

Evaluation Metric. We quantify the performance of the KVM-GAN, comparing against that of existing methods, in terms of Inception Score (IS), Fréchet Inception Distance (FID), and R-precision. Each model generated 30,000 images conditioned on the text descriptions from unseen test set for evaluation.

Dataset	GAN-INT-CLS	GAWWN	StackGAN	PPGN	AttnGAN	KVM-GAN
CUB	2.88±0.04	3.62±0.07	3.70±0.04	(-)	4.36±0.03	**4.75±0.07**
COCO	7.88±0.07	(-)	8.45±0.03	9.58±0.21	25.89±0.47	**30.49±0.57**

Table 1. Performance of inception scores for different methods on the CUB and COCO datasets.

- **Inception Score [21]**: The IS uses a pre-trained Inception v3 network [23] to compute the KL-divergence the conditional class distribution and the marginal class distribution. A large IS means that the generated model outputs a high diversity of images for all classes and each image clearly belongs to a specific class. Though the IS is widely used in recent generative models, the distance to the real-world images is still unknown.

- **Fréchet Inception Distance [5]**: The FID computes the Fréchet distance between the synthetic and real-world images based on the exracted features from Inception v3 network. A lower FID represents the generated images is closer to the real-world images.

- **R-precision**: Following Xu et al. [24], we use the R-precision to evaluate whether the generated image is well conditioned on the given text description by retrieving relevant text given a image query. In practice, we compute cosine distance between a global image vector and 100 candidate sentence vectors, which is extracted from the image and text encoders learned in DAMSM. The candidate text descriptions include one ground truth and 99 randomly selected mismatching descriptions. For each query, if the top 1 retrieval result is the ground truth, then the R-precision is 1. Otherwise, the R-precision is defined as 0. We divide the generated images into ten folds for retrieval and then take the mean and standard deviation of the resulting scores.

Implementation Details. We employ the pre-trained text and image encoding network in the DAMSM from AttnGAN and fix their parameters during training. For image generation, we adopts a similar architecture of AttnGAN, which consists of three stages where the first stage generates initial images with 64*64 resolution and next two stages refine images to 128*128 and 256*256 resolutions conditioning on word-level information using key-value memory. We apply spectral normalization [13] to all discriminator networks to avoid unusual gradients and improve text-to-image synthesis performance. All networks are trained using ADAM optimizer [6] with batch size 10 and a learning rate of 0.0002.

4.1. Text-to-image Quality

We conduct experiments to compare KVM-GAN with the state-of-the-art models on the CUB and COCO test dataset.

Dataset	Metric	AttnGAN	MemoryGAN
CUB	FID↓	23.98*	**16.09**
	R-precision↑	67.82±4.43	**72.31±0.91**
COCO	FID↓	35.49*	**32.64**
	R-precision↑	85.47±3.69	**88.56±0.28**

Table 2. Performance of FID and R-precision for different methods on the CUB and COCO datasets. The FID of AttnGAN is calculated from officially released weights.

IS. As shown in Table 4, our KVM-GAN achieves 4.80 inception score on the CUB dataset, which outperforms the inception score of AttnGAN. In addition, KVM-GAN improves the inception score from 25.89 to 30.43. The experimental results indicate that our KVM-GAN model learns a better data distribution than previous approaches.

FID. Table 4 shows the performance of FID on the CUB and COCO datasets. Our KVM-GAN decrease the FID from 23.98 to 16.00 on the CUB dataset and from 35.49 to xxxx on COCO dataset, which demenstrates that KVM-GAN generates more photo-realistic images. Beacuse the KVM-GAN is more effective to utilize word-level information using key-value memory in text-to-image task.

R-precision. As shown in Table 4, the KVM-GAN improves the R-precision about 7% on both the CUB and COCO datasets. Higher R-precision indicates that the generated images by the KVM-GAN is better conditioned on the given text description, which further demonstrates the effectiveness of the employed key-value memory.

In summary, our KVM-GAN is much superior to other text-to-image synthesis methods by a large margin.

4.2. Visual Quality

For qualitative evaluation, Figure 3 shows text-to-image synthesis examples by comparing our KVM-GAN with result of the AttnGAN. In Figure 3(a), we compare some image samples generated by the KVM-GAN and AttnGAN on the CUB dataset. Compared with AttnGAN, our KVM-GAN generates images with more vivid background and convincing details, resulting a more photo-realistic images. As shown in Figure 3(b), the imaged results seems less photo-realistic due to the difficulty of the COCO dataset. Our KVM-GAN still generates images with better quality, which is better conditioned on text descriptions.

Figure 3 shows the initial images and the refined images. In most cases, the initial images generates blurry images with rough shapes and few details. In the next stage, im-

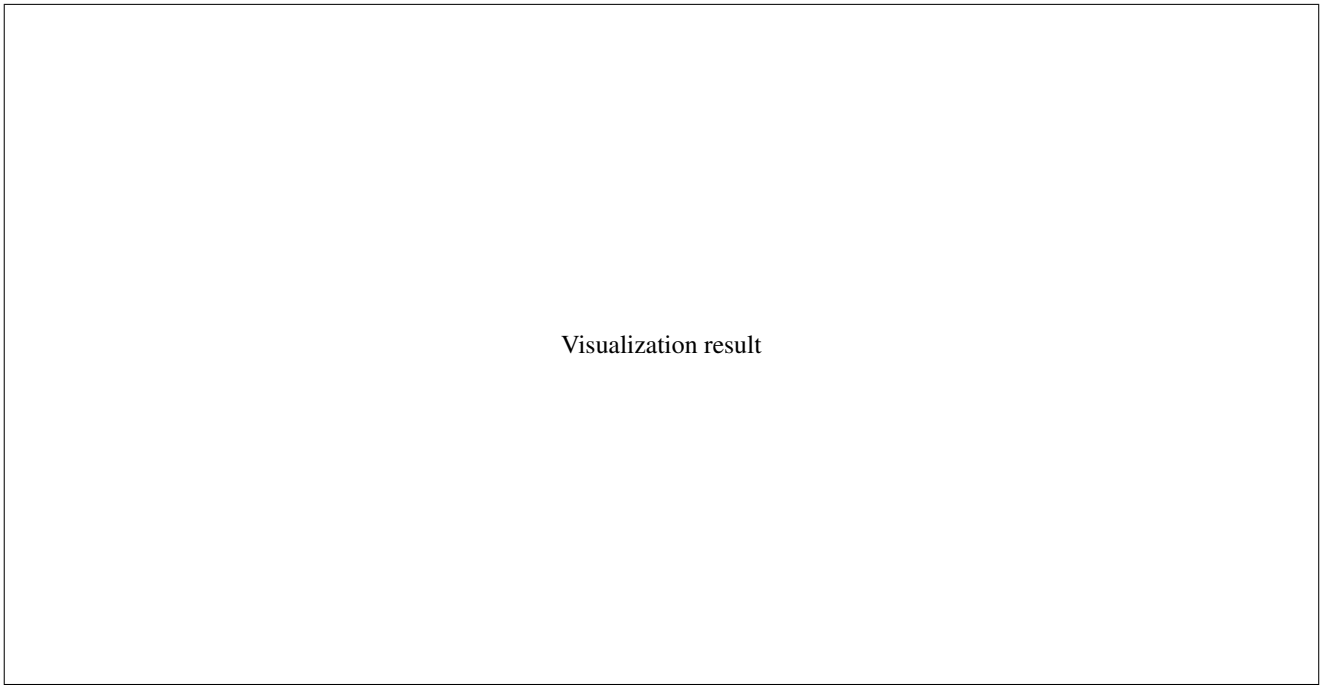

Figure 3. Example results for text-to-image synthesis by our KVM-GAN and AttnGAN. (a) Generated bird images by conditioning on text from CUB test set. (b) Generated images by conditioning on text from COCO test set.

Figure 4. The results of different stages of our KVM-GAN. The first row shows the initial images with 64 * 64 resolution. The second row shows the images with 128 * 128 resolution after one enhancement process. The third row gives the refined images with 256 * 256 resolution after two enhancement process.

ages are refined with word-level information. The refined images first enhance the foreground objects with correct color and fine-grained feathers. For instance,... Then the background is fine-tuned to be more realistic with twigs and leaves. In summary, our KVM-GAN has the ability to refine low-resolution images and generates more photo-realistic high-resulution images.

5. Conclusions

In this paper, we presente a new architecture called MemoryGAN which includes a memory component. The memory component is especially powerful to focus on im-

portant words and improve the image quality. Experiment results on two real-world datasets show that MemoryGAN outperforms the AttnGAN by both qualitative and quantitative measures. In the future, we plan to develop MemoryGAN with more powerful architectures and apply the model to harder synthesis tasks (e.g. text-to-video synthesis). A future research direction is to encode the relationship in the text description into the image. Two images with the same objects are different due to the position or interaction between objects (e.g. man riding bicycle and man pushing bicycle).

References

[1] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal. Tac-gan-text conditioned auxiliary classifier generative adversarial network. *arXiv preprint arXiv:1703.06412*, 2017. 2

[2] H. Dong, S. Yu, C. Wu, and Y. Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017. 2

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. pages 2672–2680, 2014. 1, 2

[4] C. Gulcehre, S. Chandar, K. Cho, and Y. Bengio. Dynamic neural turing machine with continuous and discrete addressing schemes. *Neural computation*, 30(4):857–884, 2018. 2

[5] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 5

[6] D. Kinga and J. B. Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, volume 5, 2015. 5

[7] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[8] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014. 2

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[10] F. Liu and J. Perez. Gated end-to-end memory networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 1–10, 2017. 2

[11] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, 2016. 1, 2

[12] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *CoRR*, 2014. 2

[13] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 5

[14] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4467–4477, 2017. 2

[15] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, pages 2642–2651, 2017. 2

[16] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 4

[17] A. Prakash, S. Zhao, S. A. Hasan, V. V. Datla, K. Lee, A. Qadir, J. Liu, and O. Farri. Condensed memory networks for clinical diagnostic inferencing. In *AAAI*, pages 3274–3280, 2017. 2

[18] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[19] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016. 2

[20] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *33rd International Conference on Machine Learning*, pages 1060–1069, 2016. 1, 2

[21] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 5

[22] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015. 2

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[24] Q. H. H. Z. Z. G. X. H. X. H. Tao Xu, Pengchuan Zhang. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 4, 5

[25] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4

[26] J. Weston, S. Chopra, and A. Bordes. Memory Networks. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 2

[27] M. Yuan and Y. Peng. Text-to-image synthesis via symmetrical distillation networks. *arXiv preprint arXiv:1808.06801*, 2018. 2

[28] H. Zhang, T. Xu, and H. Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision*, pages 5908–5916. IEEE, 2017. 2

[29] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2